Summer 7-12-2019

Evaluation of Search Engines using Advanced Search Technique: A Comparative analysis of Yahoo and Bing

Irfan ul Haq Akhoon
University of Kashmir, India, irfanakhoon@kashmiruniversity.net

Evaluation of Search Engines using Advanced Search: Comparative analysis of Yahoo and Bing

Irfan ul Haq Akhoon
Research Scholar
E-mail: irfanakhoon@kashmiruniversity.net
Ifra Rahim
MLIS Student
Humaira Mushtaq
MLIS Student
&
Dr. Shabir Ahmad
Associate Professor

Abstract

Purpose: The purpose of the paper is to evaluate the performance and efficiency of the two most used search engines after Google, i.e. Yahoo & Bing in retrieving internet resources at specific points of time using advance search techniques on single and double word queries.

Design/Methodology/Approach: The study starts from an investigation of existing methodologies for evaluating search engines in order to find out the most important factors to decide which search engine to use when searching the World Wide Web. In order to examine retrieval efficiency of a search engine on the bases of various laid parameters like, coverage of a search engine, number of dead, missing & duplicate links retrieved by a search engine by using 20 single & double word queries by adopting advanced search technique. The data were evaluated using MS Excel spread sheet software.

Findings: The findings of study reveals an explicatory results which clearly describes that different web search engines use different technology to find a particular web information. The overall analysis of the findings reveals that Yahoo is the leading search engine followed by Bing in terms of retrieving score, however Bing takes the lead in retrieving less number of dead and duplicate links while routing two term queries.

Originality/value – The paper will provide important insight into the effectiveness of two major search engines and their ability to retrieve relevant internet resources. This paper has produced key findings that are important for all web search engine users as well as researchers and the web industry. The findings will also assist search companies to improve their services.

Keywords: Search Engines, Information Retrieval, Retrieval Efficiency, Yahoo, Bing, Dead Links, Duplicate Links etc.

Introduction

The surprising growth of the web propelled the rapid development of web search engines. It has been observed in the literature that most users view search engines as the best method

available for finding information on the web. However, a user can search for any information by passing query in form of keywords or phrases. It then searches for relevant information in its database and return to the user. A large number of documents get added on the daily basis on the web and more materials become available electronically which reveals that the influence of search engines on our lives will continue to grow **(Lewandoski, 2008).** Meanwhile, the evaluation of these search engines has not been keeping up with the pace of their development. The significance of the evaluation of web search engines is twofold: to help web users in their choice of search engines and to inform the development of search algorithms and search engines. The goal of this paper is to conduct a study to measure the retrieval performance and efficiency of search engines among the two popular web search engines i.e. Yahoo! and Bing.

Nowadays, a number of search engines exist but not even a single search engine can provide a complete and inclusive expression of the web. Search engines are not able to index the whole web **(Ozcal, Altingovde & Ulusoy, 2011)**. A number of retrieval measures exist nowadays which can be used to evaluate the performance of search engines like precision, relative recall, duplicate links, dead links, unique links etc. **(Brin & Page, 1998**).

Objectives

- To select search engines and search terms for the study.
- To ascertain the coverage of selected search engine.
- To check the duplication of results indexed by each search engine.
- To identify the dead links among results indexed by each search engine.

Methodology

Multipronged methodology was adopted for the current study. Methodology adopted for the current study is presented into following subheading:

Selection of search engines

For achieving objective 1st, a thorough literature review has been carried out in order to select search engines for the study. While going through literature, it was found that among various search engines Google was always at the first place and at next it was sometimes Yahoo or some any other search engine, but Google was never at the second place. It was made understood that

evaluating Google is mere wastage of time and resources, thus 2nd positioned search engine by various studies were selected for the current study are Yahoo and Bing

Selection of search terms

Keyword were collected from the research articles which were indexed of Web of Science in Open access journals related to the field of Library and information science but only from those issues which were published in 2018. 170 keywords were collected from 76 OA research articles, as number of keywords were large, in order to get the homogeneity in keywords various keywords were dropped from the list, comprising of duplicate, complicated words and keywords of more than three terms. After elimination these keyword, ***Kerjice & Morgan*** sample calculator were used with 90% confidence and 5% of Margin of error, only 20 keywords were selected on random sampling method which includes 13 two term and 7 single term keywords.

Search Engine Relevancy

Each search engine retrieved a large number of result but ***only first twenty results*** were selected for the purpose of evaluation.

For achieving objective 2nd that is coverage of selected search engines each keyword was routed in both the two search engines to find out results.

For achieving objective 3rd, 4th retrieved results were quantified on the basis of number of duplicate links, number of dead links available among the top twenty results and were taken into consideration for further process of research.

Scope of study

The scope of the present study is confined to two know the retrieval efficiency of two selected search engines viz., Yahoo & Bing. The retrieval efficiency is measured on the basis of various set parameters like, Coverage of search engine in the form of total number of results retrieved, number of duplicate and dead links etc.

Review of Related Literature

The progress and development of any nation depends on the information because present society is information-based society. A user can have an access or retrieval of such information from every corner of world **(Brinkley & Burke, 1995**). Internet has become increasingly primary source for many users and in order to retrieve information from the web, users make use of various

tools viz., search-engines, meta-search engines etc. available on the internet **(Arampatizs, Efraimidis & Drosatos, 2013)**. Various studies had been carried out in order to evaluate the available search engines by using different evaluating parameters following are some of the recently explored studies on the said topic.

Cambazoglu and Yates (2016) studies scalability and efficiency challenges in large scale web search engines. The studies provide an in depth architectural overview of a web search engine mainly focusing on the web crawling, indexing, and query processing components. The scalability and efficiency encountered in these components are presented at four different granularities i.e.; at the level of a single computer, a cluster of computers, a single data center, and a multi-center search engine. **Ali and Gul (2016)** evaluated search engine effectiveness using query. This study reveals that as the quantity of information increases on the internet it really becomes hard for a user to retrieve the relevant information. Methodology is divided into two sections viz; selection of search engines and selection of queries. These include the fact that only few queries have been selected for this evaluation. Future research would need to include a larger and more diverse sample of queries with different levels of domain expertise and degrees of familiarity with information retrieval systems. **Sánchez, Martínez-Sanahuja & Batet (2018)** studied simulating search engines where they have found discrete event simulation to be a useful tool in this context because it enables users to both represent the actual system in a one to one correspondence with its main components and simulate the cost of their relevant operations in a precise and high level manner. This requires modeling the cost of the different operations involved in processing very large streams of user queries both at macroscopic and microscopic level**. Egri and Bayrak (2014)** find that Search engines are the biggest source of access to information on the internet and their importance is increasing day-by-day. Therefore, search engines, provides right content with in a right time, and gained so much importance nowadays. In recent years Google have provided important updates for fighting spam sites, called Panda and Penguen. When we examine these updates, we can see that there are lots of innovations and new factors for SEO. **Balbantary, Swain and Sahoo (2013)** studied that Search engines help the user to surf the web. Due to the vast number of web pages it is highly impossible for the user to retrieve the appropriate web page he needs. Thus, Web search ranking algorithms play an important role in ranking web pages so that the user could retrieve the page which is most relevant to the user's query. This paper presents a study of the applicability of two user-effort-sensitive evaluation measures on five Web search engines (Google, Ask, Yahoo, AOL and Bing). Twenty queries were collected from the list

of most hit queries in the last year from various search engines and based upon that search engines are evaluated.

Data Analysis and Interpretation

A: One Word Query

Total number of results retrieved:

In order to explore the total number of results retrieved from each search engine data collected in this regard is given in table 1.1

Table: 1.1
Number of Results Retrieved by Each Search Engine

Query No.	Search Terms	Yahoo	Bing
1.	E-Books	23, 200, 000	2, 29, 00, 000
2.	Cataloging	2, 020, 000	25, 00, 000
3.	Circulation	12, 800, 000	12, 800, 000
4.	Acquisition	13, 300, 000	32, 200, 000
5.	Indexing	14. 000, 000	95, 10, 000
6.	Journals	5, 200, 000	1, 530, 000, 000
7.	Citation	14, 900, 000	1, 47, 00, 000
Total		84, 420, 000,	400, 6100, 000

From Table 1.1, it is evident that maximum numbers of results were retrieved from ***Bing*** (400, 6100, 000) followed by ***Yahoo*** (84, 420, 000) respectively. This table shows that Bing has higher result retrieval efficiency than Yahoo.

Dead links:

The collected data regarding total number of dead links retrieved from each search engines is presented under Table 1.2.

From Table-1.2, it is retrieved by ***Bing*** (1%) is leading in retrieving more dead links as compared to yahoo, as (0%) of dead links have been found from the result retrieved from Yahoo. It indicates that Yahoo is much updated than Bing.

Table 1.2
Total no. of dead links among top 20 results

Query. No.	Search terms	Yahoo (N=140)	Bing (N=140)
1.	E-Books	0	0
2.	Cataloging	0	0
3.	Circulation	0	0
4.	Acquisition	0	1
5.	Indexing	0	0
6.	Journals	0	0
7.	Citation	0	0
Total		0 0%	1 0. 74%

Duplicate links

The collected data regarding total number of duplicate links retrieved from each search engine are given below in table 1.3

Table 1.3
Number of Duplicate links

Query. No	Search Term	Yahoo (N=140)	Bing (N=140)
1.	E-Books	2	2
2.	Cataloging	2	5
3.	Circulation	1	1
4.	Acquisition	0	1
5.	Indexing	2	2
6.	Journals	5	2
7.	Citation	2	2
Total		14 10%	15 10%

From the data of duplicate links, it is obvious that both ***Yahoo*** and ***Bing*** retrieved equal number of duplicate links (10%) each. It indicates from the table below that Yahoo and Bing shows the same number of duplicate links as they have a vast coverage.

B. Two Word Query

In this section, results are retrieved by using advanced search technique with two-word search terms from two selected search engines (Yahoo and Bing) by using various parameters. Total number of results retrieved is distributed under following sub headings.

Total Number of Results Retrieved

The number of results retrieved from each search engine is given below in table 1.4.

Table-1.4
Number of Results Retrieved from each search engine

Query No.	Search term	Yahoo	Bing
1	Academic libraries	455, 000	30, 30, 000
2	Open access	620, 000, 000	6, 34, 000
3	Digital preservation	579, 000	5, 49, 000
4	Information “services	6, 970, 000	79, 80, 000
5	Library collection	1, 080, 000	1, 51, 00, 000
6	Information seeking	592, 000	3, 58, 000
7	Electronic resources	1, 040, 000	8, 20, 000
8	Information retrieval	1, 940, 000	14, 00, 000
9	Information technology	11, 800, 000	1, 18, 00, 000
10	Information professionals	12, 800, 000	15, 000, 000
11	Library professionals	41, 900	1, 35, 000
12	Impact factor	5, 280, 000	1, 15, 00, 000
13	Library services	3, 310, 000	89, 40, 000
Total		**665887900**	**15346322**

Table-1.4 reveals that maximum numbers of results are retrieved from ***Yahoo*** (665887900) followed by ***Bing*** (15346322). On the basis of the interpretation drawn from the above table Yahoo has retrieved large no of results which means Yahoo is having vast coverage than Bing.

Dead links:

The collected data regarding total number of dead links retrieved from search engines is given below in table 1.5.

Table 1.5
No. of Dead Links Among Top 20 Results

Query No.	Search terms	Yahoo (N=260)	Bing (N=260)
01	Academic libraries	0	0
02	Open access	0	0
03	Digital preservation	0	0
04	Information “services	0	0
05	Library collection	0	0
06	Information seeking	0	0
07	Electronic resources	2	1
08	Information retrieval	3	0
09	Information technology	0	0
10	Information professionals	0	0
11	Library professionals	1	1
12	Impact factor	0	0
13	Library services	0	0
Total		**6 (2%)**	**2 (0. 76%)**

From the data of above table it is revealed that 6% of dead links where retrieved by ***Yahoo***. However (2%) is retrieved from ***Bing***. It indicates that Bing is much updated than Yahoo.

Duplicate links:

The collected data regarding total number of duplicate links retrieved from search engines is presented in table 1.6.

From the data of duplicate links it is obvious that ***Bing*** has obtained maximum number of dead links (11%), ***Yahoo*** is at the second number (9%). It indicates from the table above that there are variation among results retrieved by two search engines i. e.; Yahoo and Bing.

Table 1.6
Number of Duplicate links

Query No.	Search term	Yahoo (N=260)	Bing (N=260)
01	Academic libraries	1	0
02	Open access	0	4
03	Digital preservation	1	2
04	Information services	3	2
05	Library collection	3	4
06	Information seeking	3	3
07	Electronic resources	2	4
08	Information retrieval	1	3
09	Information technology	4	3
10	Information professionals	4	3
11	Library professionals	0	1
12	Impact factor	2	0
13	Library services	1	0
Total		**25 (9%)**	**29 (11%)**

Findings of study

The main findings of the study derived from ***Analysis and Interpretation of Data*** with particular reference to the retrieval effectiveness of each search engine using advance search for one-word and two-word queries are recorded under following two sub headings:

Advanced Search Using One Word Query:

- While accessing the ***Total number of Results*** retrieved from both the search engine, it is evident from ***Table 4.7*** that maximum numbers of results are retrieved from ***Bing*** (400, 6100, 000) followed by ***Yahoo*** (84, 420, 000). Thus, it indicates that ***Bing*** has higher result of retrieval efficiency.
- In order to estimate the ***Total number of Dead links*** retrieved from both the search engine, it is evident from ***Table 4.8*** that maximum number of dead links are retrieved from Bing (1%) and no dead link (0%) was retrieved from ***Yahoo*** and ***Bing***. Hence it indicates that ***Yahoo*** is much updated than ***Bing***.

- While calculating the ***Total number of Duplicate links*** retrieved from each search engine, it is evident from ***Table 4.9*** that ***Yahoo*** and ***Bing*** retrieved maximum number of Duplicate links.

Advanced Search Using Two Word Queries:

- In order to know the maximum numbers of ***Results Retrieved*** from both the search engine, it is apparent from ***Table 4.10*** that maximum number of results are retrieved from ***Yahoo*** (665887900) followed by ***Bing*** (15346322). Thus it is clear that ***Yahoo*** has retrieved large number of results and is having vast coverage than ***Bing***.
- In order to estimate the total number of **Dead links** retrieved from both the search engines it is evident from ***Table 4.11*** that maximum number of Dead links (2%) are retrieved from ***Yahoo***. However, there was less percentage (0. 76%) retrieved by ***Bing.*** Thus, Bing retrieved more updated results than ***Yahoo***.
- While knowing the total number of ***Duplicate links*** retrieved from both the search engines it is evident from ***Table 4.12*** that maximum numbers of duplicate links are retrieved from ***Bing*** (11%) followed by ***Yahoo*** (9%).

Comparison of Results

While comparing the result and finding from all the applied parameters of evaluating a search engine various assertions were made out of it and are listed as per the type of parameter.

- '**Number of results retrieved'** from the results retrieved out of two selected search engine by routing single & double word queries in each using ***Basic*** and ***Advanced*** search technique separately, it was found that ***Yahoo*** has retrieved highest results in **Basic** Search using **one word query, two word query** and **Advanced** search using **two word query.** However, **Bing** retrieved highest number of results while applying advanced search using **one word query**.
- **'Dead Links'** majority of **Dead links** are retrieved from ***Yahoo*** in **Basic** and **Advanced** search using **two word query** followed by ***Bing*** in **Basic** search using **one word query** and **Advanced** search using **one word query** respectively.
- **'Duplicate Links'** Maximum score of **Duplicate links** are retrieved from ***Yahoo*** in **Basic** and **Advanced** search using **one word query.** However, ***Bing*** retrieved more

duplicate links by applying ***basic search*** and ***advance search*** using **two word queries** respectively.

Conclusion

Web search engine is a system that allows users to search for information on World Wide Web (WWW). All search engines works on the mechanism present in it, usually, users enter a query into a search engine in order to retrieve specific needed information in different forms. The shortfalls witnessed by the users of search engines today is the quality of results retrieved from a search engine. The first 20 results retrieved from each search engines were taken into consideration and the selected parameters were applied on these twenty results only to measure the retrieval efficiency of a search engine. Analysed results from the study reveals that, Yahoo is having vast coverage than Bing in terms of highest number of retrieved results either one word or two word search query is used. Meanwhile, the comparative analysis of the results revealed that while search for a single term query, Bing retrieved much updated results than Yahoo, However while using two word queries Yahoo retrieves more updated results than Bing. While as,

Hence, it is concluded from the study that Yahoo is the leading search engine followed by Bing in terms of retrieving highest score of results, however Bing takes the lead in retrieving less number of dead and duplicate links while using two word query.

References

Arampatzis, A., Efraimidis, P. S & Drosatos, G. (2013). Query scrambler for research privacy on the internet. Information Retrieval, 16(6), 657-679. DOI: 10.1007/s10791-012-9212-1.

Brin, S., S & Page (1998). The Anatomy of Large scale hyper textual web search engines. Proceedings of the Seventh International Conference on World Wide Web, 30(7), 107-117. DOI: 10.1016/SO169-7552 (98)00110-x.

Brinkley, & Bruke, M (1995). Information retrieval from the internet: an evaluation of the tools. Internet Research, 5(3), 3-10. DOI: 10.1108/1066224951014595.

Lewandoski, D. (2011). The retrieval effectiveness of search engines on navigational queries. Aslib proceedings, 63(4), 354-363. DOI: 10.1108/00012531111148949.

Balbantaray, R.C; Swain, M; Sahoo, B.P (2013). Evaluation of web search engines based on ranking results and features. International journal of human computer interaction *(IJHCI)*, vol 4(3), 2013.

Chiou, L., & Tucker, C. (2017). Search Engines and Data Retention: Implications for Privacy and Antitrust

Carrión, J., Puntes, D. F., & Luque, E. (January 01, 2017). Simulating a Search Engine Service focusing on Network Performance. Procedia Computer Science, 108, 79-88.

Catena, M., Tonellotto, N., & Catena, M. (July 01, 2017).Energy-Efficient Query Processing in Web Search Engines IEEE Transactions on Knowledge and Data Engineering, 29(7), 1412-1425.

Egri, G., & Bayrak, C. (2014). The Role of Search Engine Optimization on keeping the User on the Site. Procedia Computer Science, 36, 335-342.

Fagan, J. C. (June 01, 2017). An evidence-based review of academic web search engines, 2014-2016: Implications for librarians' practice and research agenda. Information Technology and Libraries, 36(2), 7-47

Martinez-Sanahuja, L., Sanchez, D., (2016). Proceedings of 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES 2016. Evaluating the Suitability of Web Search Engines as Proxies for Knowledge Discovery from the Web. Procedia Computer Science, 96, 169-178.

Palanisamy, R. (February 21, 2013). Evaluation of Search Engines: A Conceptual Model and Research Issues. International Journal of Business and Management, 8(6)

Sánchez, D., Martínez-Sanahuja, L., & Batet, M. (2018). Survey and evaluation of web search engine hit counts as research tools in computational linguistics. Information Systems 73, 50-60.

Sánchez, D., Martínez-Sanahuja, L., & Batet, M. (2018). Survey and evaluation of web search engine hit counts as research tools in computational linguistics. Information Systems, 73, 50-60.